Braunschweigische
Wissenschaftliche Gesellschaft

Jahrbuch 2015

Sonderdruck
Seiten 111–119

J. Cramer Verlag · Braunschweig
2016

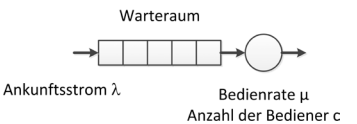
Warum steht man so oft in der falschen Warteschlange?*

Thomas Hanschke

TU Clausthal, Adolph-Roemer-Straße 2A, D-38678 Clausthal-Zellerfeld
E.Mail: hanschke@math.tu-clausthal.de

1. Das Grundmodell der Warteschlangentheorie

Die Warteschlangentheorie bedient sich zur Beschreibung von Bedienungssystemen eines einfachen Grundmodells. Es besteht aus dem sogenannten Bedienungsschalter, der über eine oder mehrere parallel arbeitende gleichartige Maschinen bzw. Arbeitsplätze verfügt, und aus einem Warteraum. Die Kunden treffen einzeln und zu zufälligen Zeitpunkten vor dem Bedienungsgerät ein. Ein neu ankommender Kunde wird bedient, sofern mindestens eines der Bedienungsgeräte frei ist, andernfalls muss er sich in die Warteschlange einreihen.



Das Grundmodell kann auf vielfältige Weise variiert werden:

– Die Kunden werden nicht einzeln, sondern gruppenweise bedient (Wartesysteme mit Gruppenbedienung). Anwendung: Losfertigung in einem Produktionsbetrieb

– Einige Kunden verlassen das System, bevor sie bedient worden sind (Wartesysteme mit Zeitbeschränkungen). Anwendung: Lagerhaltung von verderblicher Ware

– Nicht alle Bedienungsgeräte stehen jedem Kunden zur Verfügung (Bedienungssysteme mit eingeschränkter Erreichbarkeit). Anwendung: Fertigungsstraßen mit dedizierten Maschinen, Koppelanordnungen in einem Fernsprechnetz.

– Einige Kunden scheuen sich, in das Bedienungssystem einzutreten, weil ihnen die Warteschlange zu lang erscheint (Wartesystem mit ungeduldigen

---

* Der Vortrag wurde am 11.07.2015 vor der Plenarversammlung der Braunschweigischen Wissenschaftlichen Gesellschaft gehalten.

Kunden). Anwendung: übliches Kundenverhalten an einem Post-, Bank- oder Fahrkartenschalter

– Ein Kunde mit höherer Priorität verdrängt einen Kunden niedrigerer Priorität aus dem Bedienungsprozess (Bedienungssysteme mit Prioritätssteuerung). Anwendung: Express-Los-Steuerung in einem Fertigungsprozess

– Ein Kunde, der bei seiner Ankunft nicht sofort bedient werden kann, geht verloren (Verlustsysteme). Anwendung: Telefonate in einem Fernsprechnetz

Der Strom der ankommenden Forderungen wird durch einen sog. Erneuerungsprozess beschrieben. Dazu denken wir uns alle Forderungen in der Reihenfolge ihrer Ankünfte durchnummeriert. Die Zeitspanne I_n zwischen der Ankunft des (n-1)-ten und des n-ten Kunden wird als Zwischenankunftszeit bezeichnet. Von den Zufallsvariablen I_n, n = 1, 2, ... wird vorausgesetzt, dass sie stochastisch unabhängig und identisch verteilt sind mit der Verteilungsfunktion $F_I(x)$, dem Erwartungswert E[I] und der Varianz D[I]. Der Kehrwert

$$\lambda = \frac{1}{E[I]}$$

heißt Ankunftsrate und gibt an, wie viele Kunden im Durchschnitt pro Zeiteinheit in das System einfallen.

Die Bedienungszeiten S_n, $n = 1, 2, \ldots$ der aufeinanderfolgenden Kunden werden ebenfalls als stochastisch unabhängige und identisch verteile Zufallsvariablen aufgefasst. Die Verteilungsfunktion der Bedienungszeiten wird mit $F_S(x)$ bezeichnet. Für den zugehörigen Erwartungswert und die zugehörige Varianz verwenden wir die Symbole E[S] und D[S]. Der Kehrwert

$$\mu = \frac{1}{E[S]}$$

heißt Bedienrate und gibt an, wie viele Kunden im Durchschnitt pro Zeiteinheit von dem Bedienungssystem abgefertigt werden können. Sind mehrere parallele und gleichartige Bedienungsgeräte vorhanden, erhöht sich die Bedienrate entsprechend der Anzahl der Geräte.

Bedienregeln

Die Bedienungsregel legt fest, in welcher Reihenfolge die wartenden Kunden abgefertigt werden sollen. Folgende Regeln und Bezeichnungen sind gebräuchlich:

FIFO (FCFS)	First In, First Out (First Come, First Served). Die Bedienung erfolgt in der Reihenfolge der Ankünfte.

LIFO (LCFS)	Last In, First Out (Last Come, First Served). Die Bedienung erfolgt in umgekehrter Reihenfolge der Ankünfte.
SIRO	Selection In Random Order. Der nächste Kunde wird zufällig ausgewählt.
Non-preemptive priority	Relative Priorität. Manche Kunden werden gegenüber anderen Kunden vorrangig behandelt. Der laufende Bedienungsprozess wird jedoch nicht unterbrochen.
Preemptive priority	Absolute Priorität. Besitzt der neu ankommende Kunde gegenüber den anderen Kunden im System eine höhere Priorität, so wird der laufende Bedienungsprozess unterbrochen und mit der neuen Forderung fortgesetzt. Die alte Forderung wird zurückgestellt.
RR	Round Robin. Jeder Kunde kann den Bediener jeweils nur für ein bestimmtes Zeitintervall in Anspruch nehmen. Kunden, deren Abfertigung mehr Zeit benötigt, müssen sich deshalb mehrmals hintereinander in die Warteschlange einreihen.

Kendall Notation

Zur symbolischen Kennzeichnung von Bedienungssystemen haben D.G. Kendall und B.W. Gnedenko die Notation

A / B / c /m

eingeführt. Die Buchstaben **A** und **B** markieren hierbei den Verteilungstyp der Zwischenankunftszeiten und der Bedienungszeiten. Der Buchstabe c steht für die Anzahl der parallelen Bediener und m bezeichnet die Kapazität des Warteraums.

Für den Verteilungstyp sind folgende Abkürzungen gebräuchlich:

D	Deterministische Verteilung
M	Exponentialverteilung
Ek	Erlang-Verteilung mit Parameter k (k = 1, 2, ...)
Hk	Hyperexponentialverteilung mit Parameter k (k = 1, 2, ...)
PH	Phasentyp-Verteilung
G	Allgemeine Verteilung

Beispiel:

Die Notation M/G/3/5 kennzeichnet ein Bediensystem mit exponentialverteilten Zwischenankunftszeiten, beliebig verteilten Bedienzeiten, drei parallelen Bedienern und einem Warteraum, in dem maximal 5 Kunden warten können.

Stochastische Prozesse

Die Leistungsbewertung von Bedienungssystemen erfolgt auf der Basis folgender Prozesse.

- Anzahl Kunden im System $(N_t)_{t>0}$. Dieser Prozess gibt an, wie viele Kunden sich zur Zeit t im Bedienungssystem aufhalten.
- Der Prozess der aufeinanderfolgenden Verweilzeiten (bzw. Durchlaufzeiten) $(V_n)_{n \in N}$. Die Zufallsvariable V_n bezeichnet die Zeit, die der n-te Kunde im Bedienungssystem verweilt.

Zur Berechnung der Kenngrößen können verschiedene Methoden der Theorie der Stochastischen Prozesse herangezogen werden. Die Eignung einer Methode hängt sehr stark davon ab, welche Verteilungstypen für die Zwischenankunfts- und Bedienungszeiten zugrunde gelegt werden und ob zeitabhängige oder stationäre Größen berechnet werden sollen. Schon das Grundmodell der Warteschlangentheorie ist so kompliziert, dass es unter allgemeinen Verteilungsannahmen nicht exakt gelöst werden kann. Es existieren allerdings Näherungsformeln, die sich in der Praxis recht gut bewährt haben und die die stochastische Funktionsweise von Bedienungssystemen transparent machen. Nach einer Formel von Allen-Cunnen gilt für die mittlere Anzahl Kunden im stationären Fall:

$$\mathbf{E}[N] \approx \frac{\rho}{1-\rho} \times \sqrt{\rho^{c+1}} \times \left(\frac{c_I^2 + c_S^2}{2} \right) + \rho \times c$$

Hierbei bedeuten ρ die Auslastung des Systems und C_I^2 sowie C_S^2 die Variationskoeffizienten der Zwischenankunfts- und Bedienungszeiten. Die Formel lehrt uns, dass die Anzahl Kunden im System umso größer ist, je größer die Auslastung des Systems und die Variationskoeffizienten sind. Um eine geringe Warteschlange zu bekommen, muss man folglich genügend Kapazität bereitstellen oder die Variabilität des Systems gering halten.

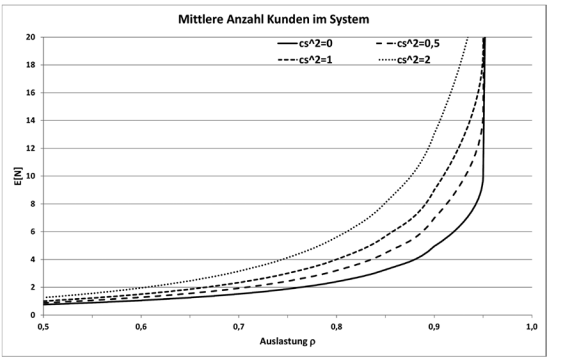
Die Mittlere Verweilzeit erhält man mit Hilfe der Formel von Little:

$$E[V] = \frac{E[N]}{\lambda}$$

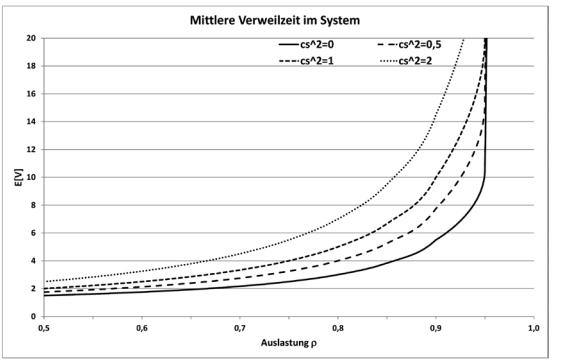
wobei λ die Inputrate des Systems bedeutet.

Ausgehend von den Formeln lassen sich folgende Zusammenhänge veranschaulichen:

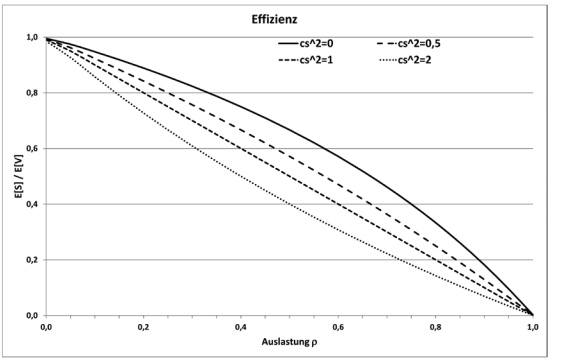
Die mittlere Anzahl Kunden im System hängt von der Auslastung der Bedienstation ab. Mit wachsender Auslastung ρ wächst auch die Anzahl Kunden im System. Außerdem beobachtet man: Je größer der Variationskoeffizient der Bedienzeit ist, umso größer ist auch die mittlere Anzahl Kunden im System E[N].



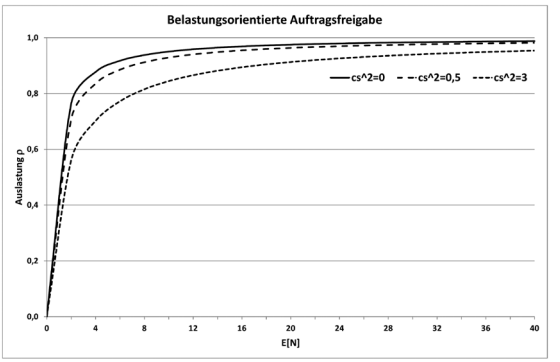
Die mittlere Verweilzeit E[V] ist offensichtlich mit der mittleren Anzahl Kunden im System korreliert. Je größer die Auslastung und die Variabilität des Systems, desto länger muss gewartet werden.



Als Effizienz bezeichnet man das Verhältnis von reiner Bedienzeit zur gesamten Verweilzeit eines Kunden. Da die Verweilzeit mit zunehmender Auslastung über alle Grenzen wächst, strebt die Effizienz gegen Null.



Unsere Formel erlaubt auch, den Auslastungsgrad ρ in Abhängigkeit vom Umlaufbestand E[N] darzustellen. Der Graph zeigt, dass man ab einem gewissen Auslastungsgrad trotz höherer Bestände keinen höheren Durchsatz erzielt. Deswegen sollten z. B. in der Produktion erst dann wieder neue Aufträge eingelastet werden, wenn der Bestand unter eine kritische Grenze gesunken ist (sogenannte belastungorientierte Auftragsfreigabe).

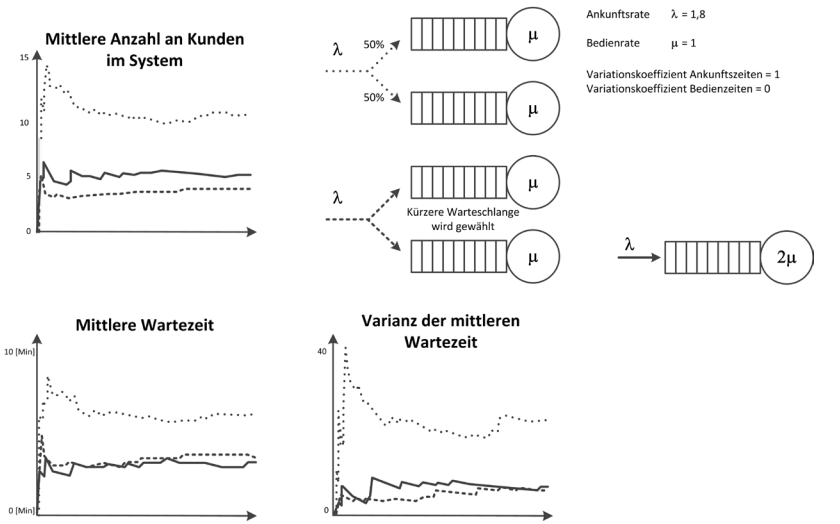


2. Warteschlangensysteme mit Steuerung

Im Folgenden wird gezeigt, dass sich Leistungsgrößen eines Warteschlangensystems schon durch geringe Steuerungsmaßnahmen des Bedienprozesses verändern bzw. verbessern lassen. Dazu werden drei verschiedene Bedienmodelle betrachtet, die sich im Einzelnen wie folgt beschreiben lassen:

a) Es stehen zwei parallele Bediener zur Verfügung. Bei der Ankunft werden die Kunden gleichmäßig (d.h. 50% pro Bediener) auf die Warteschlangen verteilt, unabhängig davon, wie viele Kunden sich bereits in den Warteschlangen befinden. Dieses Modell wird als ungesteuerter Standardfall bezeichnet.
b) Wie im Modell a) stehen auch hier zwei parallele Bediener zur Verfügung. Die Kunden werden aber nach dem Prinzip der kürzesten Warteschlange aufgeteilt, d.h. ein ankommender Kunde wird der Warteschlange mit den wenigsten wartenden Kunden zugewiesen.
c) Es gibt nur eine Warteschlange und einen Bediener. Dieser Bediener arbeitet aber mit doppelter Geschwindigkeit. Dieses Modell ist das Referenzmodell, welches im vorherigen Abschnitt am günstigsten abgeschnitten hat.

Es wurde ein Simulationslauf gestartet und so lange laufen gelassen, bis sich die verschiedenen Systeme ihrem stationären Zustand ausreichend genähert haben. Ein Vergleich im Hinblick auf mittlere Anzahl Kunden im System, mittlere Wartezeit und mittlere Varianz der Wartezeit zeigt deutlich, dass das unterlegene ungesteuerte System durch einen einfachen Steuerungsmechanismus seine Leistungsgrößen dem Referenzmodell anpassen kann und die mittlere Anzahl Kunden im System sich sogar über den Referenzwert hinaus verbessert. Die drastischen Verbesserungen



lassen sich anhand des Verlaufs der Varianz mit der Zeit leicht deuten. Durch die Verbesserungsmaßnahme entstehen gleichgroße Warteschlangen und damit gleichmäßige Wartezeiten. Die Varianz der Wartezeit nimmt deutlich ab. Beide Bediener werden im eingeschwungenen Zustand des Systems konstant zu 90% ausgelastet.

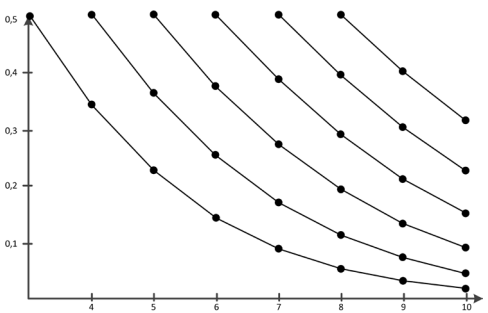
3. Warum steht man so oft in der falschen Warteschlange?

Stehen wie im Supermarkt mehrere Kassen zur Auswahl, so neigt man dazu, sich bei der Kasse mit der kürzesten Warteschlange anzustellen, um die eigene Wartezeit zu minimieren.

Doch offensichtlich garantiert diese Strategie nicht, dass man auch schneller abgefertigt wird. Da die Bediendauern der einzelnen Kunden zufällig schwanken (der eine hat mehr im Einkaufskorb, der andere weniger), kann es vorkommen, dass in der langen Warteschlange zufällig viele kleine Aufträge aufeinander folgen, während in der kürzeren Schlange große Aufträge vorherrschen. In diesem Fall muss man in der kürzeren Schlange möglicherweise länger warten als in der längeren Schlange.

Mathematisch lässt sich nachweisen, dass diese Situation umso häufiger eintritt, je unregelmäßiger die Arbeitsaufträge der einzelnen Kunden sind. (Sind umgekehrt alle Arbeitsaufträge etwa gleich groß, d.h. haben alle Kunden etwa gleichviele Gegenstände in ihrem Einkaufskorb, so wird man an der Kasse mit der kürzeren Warteschlange sicherlich auch schneller abgefertigt werden.)

Beispiel:
Um ein Zahlenbeispiel zu nennen: Unter den im Supermarkt vorherrschenden Bedingungen beträgt die Wahrscheinlichkeit, dass eine aus 8 Personen bestehende Warteschlange schneller als eine aus 5 Personen bestehende Warteschlange abgebaut wird, ungefähr 19%. Und noch extremer: Die Wahrscheinlichkeit, dass eine aus 10 Personen bestehenden Warteschlange schneller abgebaut wird als eine nur aus 5 Personen, beträgt immerhin noch ca. 9%. Dies sind vergleichsweise hohe Prozentsätze. Und so entsteht der Eindruck, so oft in der falschen Schlange zu stehen.



Auf der x-Achse ist die Anzahl der Kunden in der jeweils längeren Warteschlange und auf der y-Achse die Wahrscheinlichkeit, länger warten zu müssen, aufgetragen. Die unterste Kurve steht für „3 Kunden in der eigenen Warteschlange“ und die oberste für „8“ Kunden in der eigenen Warteschlange“. Der am weitesten links gelegene Punkt jeder Linie liegt bei einer Wahrscheinlichkeit von 50%, was auch plausibel ist: Denn wenn sich 3 Kunden in der eigenen Warteschlange und 3 Kunden in der anderen Warteschlange befinden, so beträgt die Wahrscheinlichkeit, dass man länger warten muss, gerade 50%.

Lässt man wie an den Check-In-Schaltern an den Flughäfen alle Kunden in einer gemeinsamen Warteschlange warten, kann ein solcher Eindruck nicht entstehen. Hinzu kommt, dass durch die Zusammenführung der Warteschlangen auch die Bedienungsschalter gleichmäßiger befüttert werden. Es kann z.B. nicht vorkommen, dass ein Schalter leer steht, während sich an einem anderen Schalter die Kunden stauen. Die „amerikanische“ Regel führt dazu, dass die Wartezeiten der Kunden nicht nur verkürzt sondern zusätzlich homogenisiert werden (d.h. weniger stark streuen), wodurch gleichzeitig eine größere Wartegerechtigkeit erreicht wird (siehe Abbildung).